The remaining users rarely use the system. They have logged in a few times, but for one reason or another they never become regular users of the system. Quite often this is because a lab group will settle on having one or two graduate students or post-doctoral associates become the "computer experts" of the group, and as a result, the computer use by the other people in the lab drops to an almost non-existent level. Unfortunately, an equally prevalent reason for users to stop using the GENET account is a lack of resource time. Probably the major complaint that we get from GENET users is concerning the lack of compute time and availability of the system. One account just is not enough for that many people to share, especially when it is restricted to 2 jobs at one time. We constantly remind the GENET users to use there resources wisely. We encourage them to use the BATCH system to run job in the wee hours of the morning, and we remind them to be prepared to do their work quickly when they log in to the system, but their efforts do not seem to help the problem very much.

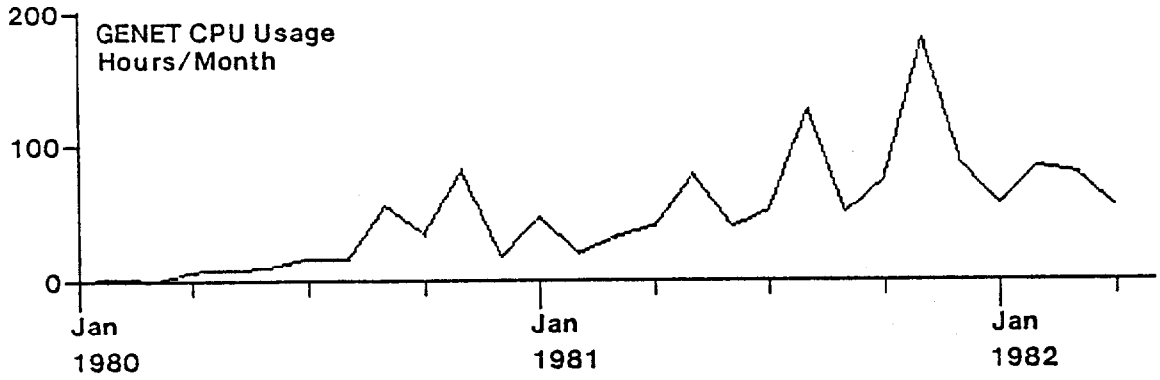
Most GENET users use only a small set of programs. These consists of text editors, which are used to set up the data files that for the MOLGEN analysis programs; XSEARCH, which GENET users use to effectively search through our database for sequences that can assist them in their research; and the electronic mail facilities. Very few of our GENET users actually feel comfortable using programs other than the ones that we maintain, not because the other programs would not be useful, but instead because the users do not have the computer time to experiment with what is available.

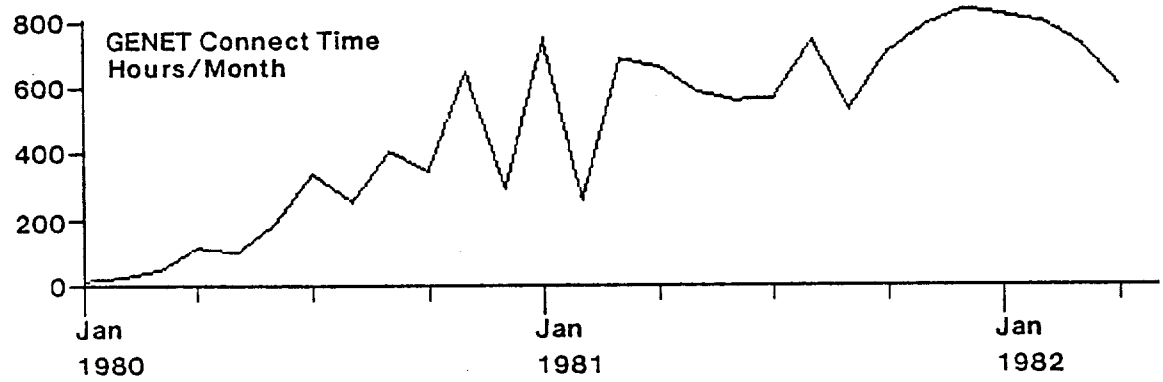
There are three note-worthy programs that we provide for GENET users that are used extensively. SEQ, a DNA-RNA sequence analysis program, is the most widely used. MAP, a program that assists in the construction of restriction maps from restriction enzyme digest data, is also used a great deal. Finally, a new program, MAPPER (written and maintained by William Pearson from Johns Hopkins University), is a simplified version of the MOLGEN MAP program that is somewhat more efficient than the MOLGEN version.

The MOLGEN UE program and special molecular genetics knowledge bases are not available to the general GENET user at this time for two reasons. First of all, the UE program is quite costly to use (in terms of computer cycles), and secondly, we feel that the knowledge base is not quite ready for the computer novice to learn and use without a significant amount of initial assistance. A few GENET users (mostly Stanford associates) that have had a significant interest in the knowledge base have become EXO-MOLGEN users and are developing knowledge bases on their own which we hope will eventually be added to the ones that MOLGEN is developing and maintaining.

GENET Usage Statistics

Following are plots of the monthly GENET CPU usage, connect time, and file usage data. The consumption of CPU time has continued to grow despite our rather stringent controls as seen in Figure 14. In fact, the cumulative GENET usage this past year is approximately 50% higher than the largest AI research project consumer as seen in Figure 11.





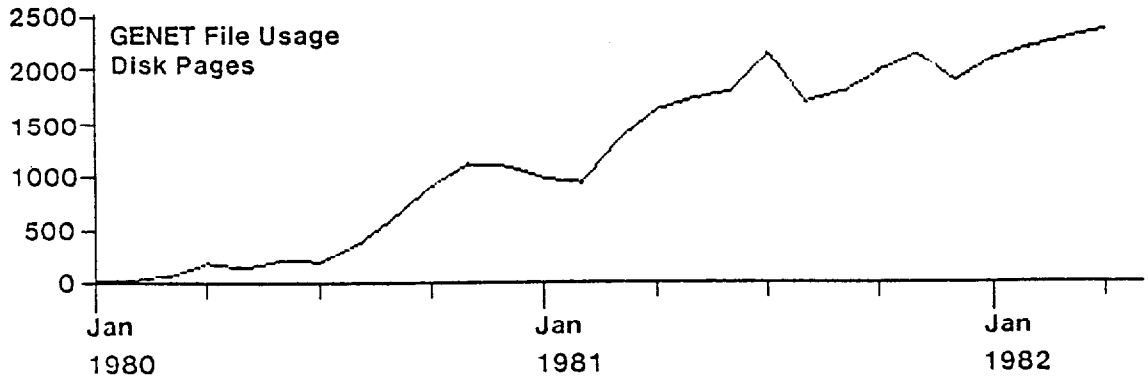


Figure 14. Monthly Resource Usage by GENET Community

I.F Comments on the Biotechnology Resources Program

Resource Organization

We continue to believe that the Biotechnology Resources Program is one of the most effective vehicles for developing and disseminating technological tools for biomedical research. The goals and methods of the program are well-designed to encourage building of the necessary multi-disciplinary groups and merging appropriate technological and medical disciplines. In our experience with the SUMEX-AIM resource, several elements of this approach seem to emerge as key to the development and management of an effective resource:

1) Effective Management Framework - there needs to be an explicit agreement between the BRP and the resource principal investigator that sets out a clear mandate for the resource and its allocation, provides worthwhile incentives for the host institution and investigator to invest the necessary substantial professional career time to develop and manage the resource, and ensures equitable distribution of resource services to its target community.

2) Close Working Relationship with NIH - a resource is a major and often long-term investment of money and human energy. A close and mutually supportive working relationship between resource management, its advisory committees, and the NIH administration is essential to assure healthy development of the resource and its relationship to its user community. We at SUMEX-AIM have benefited immensely from such a relationship with Dr. William R. Baker, Jr. in the evolution of the SUMEX-AIM community.

3) Freedom to Explore Resource Potential - a resource, by its nature, operates at the "cutting edge" in developing its characteristic technology and learning how to effectively disseminate it to the biomedical community at large. BRP should not impose artificial constraints on the resource for commercializing its efforts (fees for service) or developing its potential (funding duration limits or annual budget ceilings). Such artificial policy impositions can serve to undermine the very goals central to BRP's reason for existence. Satisfactory policies in this regard have been worked out recently and should be retained.

Electronic Communications

SUMEX-AIM has pioneered in developing more effective methods for facilitating scientific communication. Whereas face to face contacts continue to play a key role, in the longer term we feel that computer-based communications will become increasingly important to NIH and the biomedical community. We would like to see BRP take a more active role in promoting these tools within NIH and its grantee community. A concrete step would be to become a sponsoring agency for the ARPANET which remains the most effective means for a very broad spectrum of services to promote good

communications. This could serve as a base for interconnecting sponsored machines and offering a broader range of services and promoting broader collaboration among the biomedical community at large.

II Description of Scientific Subprojects

II.A Scientific Subprojects

The following subsections report on the AIM community of projects and "pilot" efforts including local and national users of the SUMEX-AIM facility at Stanford. Those using the Rutgers-AIM facility are annotated with "[Rutgers-AIM]". In addition to these detailed progress reports, we have included briefer summary abstracts of the fully authorized projects in Appendix C on page 277.

The collaborative project reports and comments are the result of a solicitation for contributions sent to each of the project Principal Investigators requesting the following information:

I. SUMMARY OF RESEARCH PROGRAM
- A. Project rationale
- B. Medical relevance and collaboration
- C. Highlights of research progress
  - --Accomplishments this past year
  - --Research in progress
- D. List of relevant publications
- E. Funding support (see details below)

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE
- A. Medical collaborations and program dissemination via SUMEX
- B. Sharing and interactions with other SUMEX-AIM projects (via computing facilities, workshops, personal contacts, etc.)
- C. Critique of resource management (community facilitation, computer services, communications services, capacity, etc.)

III. RESEARCH PLANS (8/80-7/86)
- A. Project goals and plans
  - --Near-term
  - --Long-range
- B. Justification and requirements for continued SUMEX use
- C. Needs and plans for other computing resources beyond SUMEX-AIM
- D. Recommendations for future community and resource development

We believe that the reports of the individual projects speak for themselves as rationales for participation; in any case the reports are recorded as submitted and are the responsibility of the indicated project leaders.

II.A.1 Stanford Projects

The following group of projects is formally approved for access to the Stanford aliquot of the SUMEX-AIM resource. Their access is based on review by the Stanford Advisory Group and approval by Professor Feigenbaum as Principal Investigator.

II.A.1.1 AGE - Attempt to Generalize

AGE - Attempt to Generalize

H. Penny Nii and Edward A. Feigenbaum
Computer Science Department
Stanford University

ABSTRACT: Isolate inference, control, and representation techniques from previous knowledge-based programs; reprogram them for domain independence; write an interface that will help a user understand what the package offers and how to use the modules; and make the package available to other members of the AIM community and labs doing knowledge-based programs development, and the general scientific community.

I. SUMMARY OF RESEARCH PROGRAM

A. Project Rationale

The general goal of the AGE project is to demystify and make explicit the art of knowledge engineering. It is an attempt to formulate the knowledge that knowledge engineers use in constructing knowledge-based programs and put it at the disposal of others in the form of a software laboratory.

The design and implementation of the AGE program is based primarily on the experience gained in building knowledge-based programs at the Stanford Heuristic Programming Project in the last decade. The programs that have been, or are being, built are: DENDRAL, meta-DENDRAL, MYCIN, HASP, AM, MOLGEN, CRYSALIS [Feigenbaum 1977, 1980], and SACON [Bennett 1978]. Initially, the AGE program will embody artificial intelligence methods and techniques used in these programs. However, the long-range aspiration is to integrate those developed at other AI laboratories. The final product is to be a collection of building-block programs combined with an "intelligent front-end" that will assist the user in constructing knowledge-based programs. It is hoped that AGE will speed up the process of building knowledge-based programs and facilitate the dissemination of AI techniques by: (1) packaging common AI software tools so that they need not be reprogrammed for every problem; and (2) helping people who are not knowledge engineering specialists write knowledge-based programs.

B. Medical Relevance and Collaboration

AGE is relevant to the SUMEX-AIM Community in two ways: as a vehicle for disseminating cumulated knowledge about the methodologies of knowledge engineering and as a tool for reducing the amount of time needed to develop knowledge-based programs.

(1). Dissemination of Knowledge: The primary strategy for conducting AI research at the Stanford Heuristic Programming Project is to build

complex programs to solve carefully chosen problems and to allow the problems to condition the choice of scientific paths to be explored. The historical context in which this methodology arose and summaries of the programs that have been built over the last decade at HPP are discussed in [Feigenbaum 1977, 1980]. While the programs serve as case studies in building a field of "knowledge engineering," they also contribute to a cumulation of theory in representation and control paradigms and of methods in the construction of knowledge-based programs.

The cumulation and concomitant dissemination of theory occur through scientific papers. Over the past decade we have also cumulated and disseminated methodological knowledge. In Computer Science, one effective method of disseminating knowledge is in the form of software packages. Statistical packages, though not related to AI, are one such example of software packages containing cumulated knowledge. AGE is an attempt to make yesterday's "experimental technique" into tomorrow's "tool" in the field of knowledge engineering.

(2). Speeding up the Process of Building Knowledge-based Programs: Many of the programs built at HPP are intelligent agents to assist human problem solving in tasks of significance to medicine and biology (see separate sections for discussions of work and relevance). Without exception the programs were handcrafted. This process often takes many years, both for the AI scientists and for the experts in the field of collaboration.

AGE will reduce this time by providing a set of preprogrammed inference mechanisms and representational forms that can be used for a variety of tasks. Close collaboration is still necessary to provide the knowledge base, but the system design and programming time of the AI scientists can be significantly reduced. Since knowledge engineering is an empirical science, in which many programming experiments are conducted before programs suitable for a task are produced, reducing the programming and experimenting time would significantly reduce the time required to build knowledge-based programs.

C. Highlights of Research Summary

The plans made in 1976 for the AGE project included the construction of two systems. The development of the first of these systems, AGE-1, was officially concluded on October 31, 1981. The system, together with documentation, is now available for use.

Much of the year was spent in activities related to releasing AGE-1. The most time-consuming activity was finishing the documentation. Most of the knowledge specification editors and debugging facilities were rewritten to improve the user interface. In addition, several new features were added in the area of data input protocols and focussing mechanisms.

The current user interface is directed at teletype-like terminals; that is, terminals where information is presented linearly in a single window. For a complex system that has many inter-related components that

need to be specified and manipulated (such as AGE-1), this mode of interaction makes the system appear more complex than it is. We rewrote most of the user interface to alleviate many of the problem encountered by users in the past, but there are many problems that cannot be solved without moving to another medium of communication. With this motivation, we began our experiment in using multiple windows on the bit-map display of the Dolphin. The version of AGE-1 on the Dolphin is called AGE-1.5 -- the only difference from AGE-1 is the user interface protocals. (It should be noted that moving AGE, which was optimized for a time-sharing system, to a personal computer took several weeks.)

Our plans to begin the design of AGE-2 was postponed until 1982 (see Future Research section).

D. Publications

Nii, H. Penny and Aiello, Nelleke, "AGE: a knowledge-based program for building knowledge-based programs," Proc. of IJCAI-6, pp. 645-655, vol. 2, 1979.

Nii, H. Penny, "An Introduction to Knowledge Engineering, Blackboard Model, and AGE," HPP Working Paper, HPP-80-29.

Aiello, N. and Nii, H.P., "The Joy of AGE-ing: A User's Guide to AGE-1," October 31, 1981.

Aiello, N., Bock, C., Nii, H.P., White, W., "AGE Reference Manual," October 31, 1981.

AGE Example Series 1: "BOWL: A Beginner's Program."

AGE Example Series 2: "AGEPUFF: A Simple Event-Driven Program."

II. INTERACTION WITH THE SUMEX-AIM RESOURCES

AGE Availability

Currently AGE-1 is available on the PDP-10 at the SUMEX-AIM Computing Facility and on the PDP-20/60 at the SCORE Facility of the Computer Science Department. A tape of the compiled system that will run with Tenex or Tops-20 operating systems is available for a taping fee. The current implementation is described briefly in a later section.

Summary Description of AGE-1

Currently Implemented Tools: AGE-1 provides the user with a set of preprogrammed modules called "components". Using different combinations of these program modules, the user can build a variety of programs that display different problem-solving behavior. AGE-1 also provides a user interface modules can help the user in constructing and specifying the details of the components. A component is a collection of functions and

variables that support conceptual entities in program form. For example, production rule, as a component, consists of: (1) a rule interpreter that support the syntactic and semantic description of production-rule representation as defined in AGE, and (2) various strategies for rule selection and execution.

The components in AGE-1 have been carefully selected and modularly programmed to be useable in combinations. For those users not familiar enough to experiment with combining the components, AGE-1 provides the user two predefined configuration of components -- each configuration is called a "framework". One framework, called the Blackboard framework, is for building programs that are based on the Blackboard model [Lesser 77]. Blackboard model uses the concepts of a globally accessible data structure called a "blackboard", and independent sources of knowledge which cooperate to form hypotheses. The Blackboard model has been modified to allow flexibility in representation, selection, and utilization of knowledge. The other framework, called the Backchain framework, is for building programs that use backward-chained production rules as its primary mechanism of generating inferences.

The Front-End: To support the user in the selection, specification, and use of the components, AGE-1 is organized around four major subsystems that interact in various ways. Surrounding it is a system executive that allows the user access to these subsystems, as well as other user facilities, through menu selection. Figure 1. shows the general interrelationship among these subsystems.

The Design subsystems helps to familiarize the user with AGE-1 and to guide the user in the construction of his programs through the use of predefined frameworks. The second subsystem is a collection of interface modules that help the user specify the various components of the framework. The other subsystems are designed for testing and refining the user program. Each of the subsystem is described in more detail below:

DESIGN: The function of the DESIGN subsystem is to guide the user in the design and construction of his program through the use of predefined configuration of components, or framework. Each framework is defined in DESIGN-SCHEMA, a data structure in the form of AND/OR tree, that, on one hand, represents all the possible configuration of components within the framework; and, on the other hand, represents the decisions the user must make in order to design the details of the user program. Using this schema, the DESIGN subsystem guides the user from one design decision point to another. At each decision point, the user has access to the "help" file and also to advice regarding design decisions at that point. An appropriate ACQUISITION module can be invoked from the DESIGN subsystem so that general design and implementation specifications can be accomplished simultaneously.

ACQUISITION: For each component that the user must specify, there is a corresponding specification editor module that queries the user for task-specific information. The calling sequence of the acquisition module is guided by DESIGN-SCHEMA when the user is using the DESIGN subsystem. They can also be accessed directly from the system executive or Interlisp.

INTERPRETER: This subsystem contains several modules that help the user run and debug his program. The Check module checks for the completeness and correctness of the specification for an entire framework. For any error found, the system can call an appropriate editor to fix the error. The Interpreter executes the user program. The Trace and Break modules are run-time debugging aids. The Editor, Check, Trace, Break, and the Explanation (described below) modules are designed to complement each other, and to help the user observe the workings of his program and to make corrections as necessary.

EXPLANATION: AGE-1 has enough information to replay its execution steps, and it has reasonable justifications for the actions taken within the various frameworks. AGE-1 provides a trace-back explanation facility whereby questions related to the execution history can be answered by the system interactively. However, AGE is totally ignorant of the user's task domain and has no means of conducting a dialogue about the specifics of the domain. A detailed history of the execution steps is available to the user to build his own domain specific explanation, if necessary.

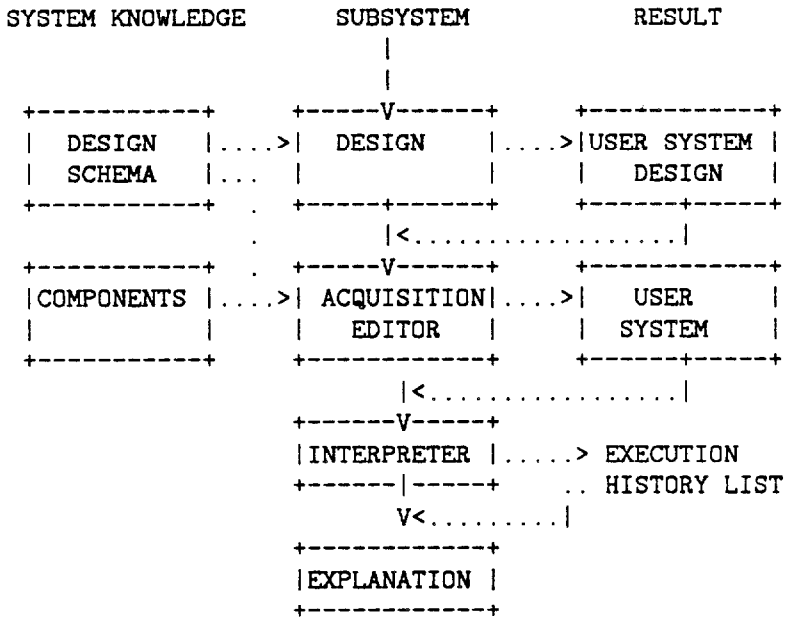


Figure 1. AGE System Organization
(... = data flow; --- = control flow)

III. RESEARCH PLAN

The primary objective of the AGE Project was, and continues to be, to see if a software laboratory could be built to speed up the process of building Expert Systems. This task was subdivided into two major subtasks:

--tool building: to isolate inference, control, and representation techniques used in other Expert Systems and reprogram them for domain independence; and

--user interface: to build an intelligent front-end to guide the user in the use of the tools.

The strategy for the tool-building task was to take paradigms with a history of successful applications, decompose them into more or less independent parts, and reprogram them. The first paradigm to be thus decomposed and reimplemented was the Blackboard model as used in HASP and CRYSALIS. Currently, AGE-1 contains components for building programs that use a rule-based blackboard model, backward-chained rules, Units, and any combination of the three. In each decomposed components we tried to extend and/or generalize; for example, in AGE-1, the user can define separate methods to deal with uncertainty for different kinds of knowledge.

The task of intelligent front-end for AGE was further broken down into two stages based on different types of users:

--Stage 1 Task: Build a system usable by an AI scientist or knowledge engineer who knows Lisp and the production rule representation of knowledge; who is familiar with methods of building knowledge-based systems; and who wants to use AGE to avoid coding basic system components and to try different problem-solving techniques with minimum recoding.

--Stage 2 Task: Build a system for a person who has a good working knowledge of AI, but who is not familiar with building Expert Systems and needs guidance on what techniques to use.

AGE-1 is a Stage 1 system. AGE-2 is to be a system directed at novice knowledge engineers. Determining the shape of AGE-2 and implementing it involves a variety of research tasks described below.

FUTURE WORK

There are many difficulties encountered by the users of AGE-1. Ninety per cent of the difficulties can be attributed to complexity of the system. It contains many design options, the consequences of which the average user does not understand, nor needs to understand for many of the applications. (It should be noted that most people interested in systems like AGE at this point in time are novice knowledge engineers.) This difficulty is compounded by the need to specify and view many interrelated parts of the user program in a linear presentation.

Solving the first problem involves research in the design aspects of Expert Systems. Although AGE-1 can be used to design many different kinds of Expert Systems, the current Design module is minimal. It is minimal in two aspects: (1) it only keeps track of parts that need to be specified and what has been specified, with some suggestion on what part to work on next; and (2) it only knows about the design of programs based on blackboard

model and that of backward chained rules. This Design module needs to be replaced by one that can help the user match his problem characteristics with appropriate combination of AGE component codes and concepts. This leads to a difficult task of doing knowledge engineering on knowledge engineers. To do this, the current components will have to be re-represented in a uniform manner, and rules written to match aspects of user problem with AGE facilities. The immediate bottleneck is in representing parts of AGE which involve the description of interrelated processes.

The more immediately doable task is to replace the linear aspect of program specification (including knowledge acquisition) by using the graphic facilities available on the Dolphins. Before any changes can be made to AGE-1, it must be brought up on a Dolphin and various system maintenance facilities implemented. This is currently being done. The system on Dolphin with the AGE-1 internals and a multi-dimensional user interface will be called AGE-1.5.

AGE-1.5

In the current AGE System, as in most other expert system building tools, knowledge is acquired from the user serially. The user is asked questions and types in the answers. In general the user works on one part of his system at a time, and must exit the current editor or acquisition module to examine or look at another part of his system.

The questions in AGE are friendly and mostly self-explanatory, requiring minimum intervention or aid from the knowledge engineer. However the serial nature of the questions does cause a bottleneck. For experienced users the questions and prompts frequently become annoying and seem to get in the way of productive work. This is true in spite of the fact that the experienced user sees an abbreviated version of explanations, comments, questions, and prompts.

To ease this problem and speed up the knowledge acquisition process, our plan is to add graphic capabilities to AGE. Using Interlisp-D on a Dolphin, we will implement menu selection, windows and screen editors, and possibly, graphic display of blackboard contents. For information which can best be acquired using serial questions, menu selection will allow the user to select the proper answer with the touch of a button. (This is much faster than typing in enough of the correct response to ambiguate it from other possible answers.) We intend to display windows showing some of each component of the user's system. Again, by a touch of a (mouse) button, the user will be able to scroll a particular window to look at specific information, if it is not already visible. These windows will also be available to trace the execution of the user's program, with the relevant, changing information in each component visible in the window at any particular time. Finally, a screen editor will be implemented to allow the user to edit information in a window or move information between windows.

The window package in AGE will be designed to allow for experimentation with the sizes and locations of various windows. Tests will be conducted to compare AGE-1 and the graphic AGE-1.5 to show that the

graphics capabilities significantly improve the knowledge acquisition process. By improve we mean both to shorten the duration of the acquisition process and to improve its palatability. Similar experiments will be used to determine the most efficient layout of the windows in the graphic AGE system.

AGE-2

AGE-2 will try to address the second of the research tasks described above.

Although the current Design subsystem provides specification functions that allow the user to interactively specify the knowledge of the domain and the control structure, it does not (aside from simple advise) provide the user any help in the actual design process. For example, AGE should be able to provide some aids to the user on what kinds of inference mechanisms and representations are appropriate for his application problem. We have stated this problem in our previous reports without any promising ideas on how we might attack this problem. With the variety of feedbacks we received from our experimental users, we now understand a few of the problems the inexperienced users are faced with. With these in mind, we have begun, and will continue, to explore ways in which we can redesign and add facilities that will help users who are not familiar with knowledge engineering techniques and methodologies.

One of the major obstacles in the way of AGE-2 development is the way in which AGE-1 is implemented. Although the syntax of AGE-1 is clearly defined (see the Reference Manual), the semantics are not well-defined. They are defined in ad hoc fashion in the Editor, the Interpreter, and the Check modules. In order for AGE-2 to be able to conduct a dialogue about itself with the user, its semantics, as well as its syntax, must be uniformly represented. Since very little research results are available in the area of representing the semantics of systems (one exception is in the automatic programming research), we need to experiment with a variety of approaches. We have already begun to look into some alternative representations. In changing the representation of the AGE system, no new components will be added, and minimum amount of changes will be made to the definition of the existing components.

Concurrent with re-representing the AGE system, we will identify a dozen or so framework, in addition of the existing two, that have simpler constructs and are easier for the novice users to understand. The simplicity will be achieved by providing less options for the user -- options which, because of their nature, are confusing to new users. Limiting the degrees of freedom for the user has the side benefit of allowing AGE to provide more specific description and aids. For example, in a very constrained framework we can provide a library of "standard" predicates for the users, which can have associated with them English translations; with such texts available the rules and the back-trace explanation can be printed in English-like form. Once the user is comfortable with the more simple frameworks, he can add complexity simply by replacing the predefined options selected for the frameworks.

AGE-2 design will not begin until AGE-1.5 is further along and until we have more data points on how AGE-1 and AGE-1.5 are used.

Computing Resources and Management

We believe the computing and communication resources provide by the SUMEX Facility is one of the best in the country. The management is responsive to the needs of the research community and provides superb services. However, the system is getting to a point where no serious research and development is possible, because of the lack of computing cycles due to overcrowding. It is a compliment to the facility that there are so many users. On the other hand, our productivity has gone down in recent months, because of the heavy load on the system. It would appear that the situation will not improve on its own, since many of the projects that were small a few years ago are maturing into larger, more complex systems. Which is the way it should be. The environment in which the work is done also needs to grow. In short, without augmentation to the current computing power and storage space (which had never been generous), our ability to make research progress at SUMEX will be drastically curtailed.

II.A.1.2 AI Handbook Project

Handbook of Artificial Intelligence

E.A. Feigenbaum, A. Barr, and P. Cohen
Stanford Computer Science Department

I. SUMMARY OF RESEARCH PROGRAM

A. Technical Goals

The AI Handbook is a compendium of knowledge about the field of Artificial Intelligence. It has been edited by Avron Barr, Paul Cohen, and Edward Feigenbaum, with textual contributions from students and investigators at several research facilities across the nation. The scope of the work is broad: Hundreds of articles cover most of the important ideas, techniques, and systems developed during 26 years of research in AI. Each short article is a description written for non-AI specialists and students of AI. Additional articles serve as Overviews, which discuss the various approaches within a subfield, the issues, and the problems.

There is no comparable resource for AI researchers and other scientists and technologists who need access to descriptions of AI techniques and concepts. The research literature in AI is not very accessible. And the elementary textbooks are not nearly broad enough in scope to be useful to a scientist working primarily in another discipline who wants to do something requiring knowledge of AI. Furthermore, we feel that some of the Overview articles are the best critical discussions of activity in the field available anywhere.

To indicate the scope of the Handbook, we have included an outline of the articles as an appendix to this report (see page 269).

B. Medical Relevance and Collaboration

The AI Handbook Project was undertaken as a core activity by SUMEX in the spirit of community building that is the fundamental concern of the facility. We feel that the organization and propagation of this kind of information to the AIM community, as well as to other fields where AI is being applied, is a valuable service that we are uniquely qualified to support.

C. Progress Summary

The major work of this project is now finished. The Handbook material was completed in April, 1982, and has been published in three volumes--over 1500 pages. The chapters also are appearing as Stanford Computer Science Department Technical Reports available through the National Technical Information Service. Work continues on developing a convenient mechanism for on-line access to the Handbook material. When

that access software is completed, the Handbook text will be available for browsing by the SUMEX community.

Both the first and second volumes of the Handbook have been selected by the Library of Science Book Club as main selections.

D. List of Relevant Publications

"The Handbook of Artificial Intelligence, Volume I," Avron Barr and Edward A. Feigenbaum, Eds., William Kaufmann, Inc., Los Altos, California, May 1981.

"The Handbook of Artificial Intelligence, Volume II," Avron Barr and Edward A. Feigenbaum, Eds., William Kaufmann, Inc., Los Altos, California, June 1982.

"The Handbook of Artificial Intelligence, Volume III," Paul Cohen and Edward A. Feigenbaum, Eds., William Kaufmann, Inc., Los Altos, California, June 1982.

Many of the chapters of Volumes I and II of the AI Handbook have already appeared in preliminary form as Stanford Computer Science Technical Reports, authored by the respective chapter-editors. References follow. Chapters from Volume III will appear as Technical Reports in the summer and fall of 1982.

HPP-79-12 (STAN-CS-79-726)
Ann Gardner. Search.

HPP-79-17 (STAN-CS-79-749)
William Clancey, James Bennett, and Paul Cohen.
Applications-oriented AI Research: Education.

HPP-79-21 (STAN-CS-79-754)
Anne Gardner, James Davidson, and Terry Winograd.
Natural Language Understanding.

HPP-79-22 (STAN-CS-79-756)
James S. Bennett, Bruce G. Buchanan, and Paul R. Cohen.
Applications-oriented AI Research: Science and Mathematics.

HPP-79-23 (STAN-CS-79-757)
Victor Ciesielski, James S. Bennett, and Paul R. Cohen.
Applications-oriented AI Research: Medicine.

HPP-79-24 (STAN-CS-79-758)
Robert Elschlager and Jorge Phillips. Automatic Programming.

HPP-80-3 (STAN-CS-80-793)
Avron Barr and James Davidson. Representation of Knowledge.

E. Funding Support Status

The Handbook Project is partially supported under the Heuristic Programming Project contract with the Advanced Research Projects Agency of the DOD, contract number MDA903-80-C-0107, E. A. Feigenbaum, Principal Investigator and under the core research activities of the SUMEX-AIM resource.

II. INTERACTIONS WITH SUMEX-AIM RESOURCE

A. Collaborations and Medical Use of Programs via SUMEX

We have had a modest level of collaboration with a group of students and staff at the Rutgers resource, as well as occasional collaboration with individuals at other ARPA net sites.

B. Sharing and Interactions with Other SUMEX-AIM Projects

As described above, we have had moderate levels of interaction with other members of the SUMEX-AIM community, in the form of writing and reviewing Handbook material. During the development of this material, arrangements were made for sharing the emerging text. The published material will also be made available to the community as an on-line resource.

C. Critique of Resource Management

Our requests of the SUMEX management and systems staff, requests for additional file space, directories, systems support, or program changes, have been answered promptly, courteously and competently, on every occasion.

III. RESEARCH PLANS (8/80 - 7/83)

A. Long Range Project Goals

During 1982, all material will be published. The on-line access program will continue under development.

B. Justifications and Requirements for Continued SUMEX Use

The AI Handbook Project is a good example of community collaboration using the SUMEX-AIM communication facilities to prepare, review, and disseminate this reference work on AI techniques. The Handbook articles currently exist as computer files at the SUMEX facility. All of our authors and reviewers have had access to these files via the network facilities and have used the document-editing and formatting programs available at SUMEX. This relatively small investment of resources has resulted in what we feel is a seminal publication in the field of AI, of particular value to researchers who want quick access to AI ideas and techniques for application in other areas.

C. Needs and Plans for Other Computational Resources

We use document preparation programs at SUMEX and the Computer Science Department's SCORE machine. We have used and will continue to use a Computer Science Department phototypesetting machine, the Alphatype, to produce the final copy of the AI Handbook. The phototypesetting software called TEX, developed at Stanford, is the vehicle for this production.

The on-line access program will be written as a SUMEX systems resource.

D. Recommendations for Future Community and Resource Development

None.

II.A.1.3 DENDRAL Project

The DENDRAL Project
Resource-Related Research: Computers in Chemistry

Prof. Carl Djerassi
Department of Chemistry
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

The DENDRAL Project is a resource-related research project. The resource to which it is related is SUMEX-AIM, which provides DENDRAL its sole computational resource for program development and dissemination to the biomedical community.

A. Project Rationale

The DENDRAL project is concerned with the application of state-of-the-art computational techniques to several aspects of structural chemistry. The overall goals of our research are to develop and apply computational techniques to the procedures of structural analysis of known and unknown organic compounds based on structural information obtained from physical and chemical methods and to place these techniques in the hands of a wide community of collaborators to help them solve questions of structure of important biomolecules. These techniques are embodied in interactive computer programs which place structural analysis under the complete control of the scientist working on his or her own structural problem. Thus, we stress the word assisted when we characterize our research effort as computer-assisted structure elucidation or analysis.

Our principal objective is to extend our existing techniques for computer assistance in the representation and manipulation of chemical structures along two complementary, interdigitated lines. We are developing a comprehensive, interactive system to assist scientists in all phases of structural analysis (SASES, or Semi-Automated Structure Elucidation System) from data interpretation through structure generation to data prediction. This system will act as a computer-based laboratory in which complex structural questions can be posed and answered quickly, thereby conserving time and sample. In a complementary effort we are extending our techniques from the current emphasis on topological, or constitutional, representations of structure to detailed treatment of conformational and configurational stereochemical aspects of structure.

By meeting our objectives we will fill in the "missing link" in computer assistance in structural analysis. Our capabilities for structural analysis based on the three-dimensional nature of molecules is an absolute necessity for relating structural characteristics of molecules to their observed biological, chemical or spectroscopic behavior. These

capabilities will represent a quantum leap beyond our current techniques and open new vistas in applications of our programs, both of which will attract new applications among a broad community of structural chemists and biochemists who will have access to our techniques. This access depends entirely on our access to and the continued availability of SUMEX-AIM. These issues are discussed in detail in the subsequent section, Interactions with the SUMEX-AIM Resource.

The primary rationale for our research effort is that structure determination of unknown structures and the relationship of known structures to observed spectroscopic or biological activity are complex and time-consuming tasks. We know from past experience that computer programs can complement the biochemist's knowledge and reasoning power, thereby acting as valuable assistants in solving important biomedical problems. By meeting our objectives we feel strongly that our programs will become essential tools in the repertoire of techniques available to the structural biochemist.

We are currently beginning the third year of our three year grant. This period represents a transition in the sense that we have pushed our research efforts in techniques for spectral interpretation, structure generation (e.g., CONGEN) and spectral prediction to their limits within the confines of topological representations of molecular structure. At this time, these techniques are perceived to be of significant utility in the scientific community as evidenced by our workshops, the demand for the exportable version of CONGEN and the number of persons requesting collaborative or guest access to our programs at Stanford (see Interactions with the SUMEX-AIM Resource). These existing techniques will, for some years to come, remain as important first steps in solving structural problems. However, in order to anticipate the future needs of the community for programs which are more generally applicable to biological structure problems and more easily accessible we must address squarely the limitations inherent in existing approaches and search for ways to solve them. Our major objectives are based on the following rationale.

None of our techniques (or the techniques of any other investigators) for computer-assisted structure elucidation of unknown molecular structures make full use of stereochemical information. As existing programs were being developed this limitation was less important. The first step in many structure determinations is to establish the constitution of the structure, or the topological structure, and that is what CONGEN, for example, was designed to accomplish. However, most spectroscopic behavior and certainly most biological activities of molecules are due to their three-dimensional nature. For example, some programs for prediction of the number of resonances observed in 13CMR spectra use the topological symmetry group of a molecule for prediction. However, in reality it is the symmetry group of the stereoisomer that must be used. This group reflects the usually lower symmetry of molecules possessing chiral centers and which generally exist in fewer than the total possible number of conformations. This will increase the number of carbon resonances observed over that predicted by the topological symmetry group alone. More generally, few of the techniques in the area of computer-assisted structure elucidation can be used in

accurate prediction of structure/property relationships, whether the properties be spectral resonances or biological activities.

A structure is not, in fact, considered to be established until its configuration, at least, has been determined. Its conformational behavior may then be important to determine its spectroscopic or biological behavior. For these reasons we are emphasizing in our current grant period development of stereochemical extensions to CONGEN, our newly-developed structure generator, GENOA (see References 17, 18), and related programs such as the C-13 Nuclear Magnetic Resonance (NMR) programs (see References 15, 16,19-23), including machine representations and manipulations of configuration (see References 1, 10) and conformation (see Reference 19,24,26) and constrained generators for both aspects of stereochemistry (see References 6, 9, 11, 12).

None of the existing techniques for computer-assisted structure elucidation of unknown molecules, excepting very recent developments in our own laboratory, are capable of structure generation based on inferred partial structures which may overlap to any extent. Such a capability is a critical element in a computer-based system, such as we propose, for automated inference of substructures and subsequent structure generation based on what is frequently highly redundant structural information including many overlapping part structures. Important elements of our research are concerned with further developments of such a capability for structure generation (the GENOA program, (see Reference 17)).

Given the above tools for structure representation and generation, we can consider new interpretive and predictive techniques for relating spectroscopic data (or other properties) to molecular structure (see References 2, 3, 7, 8, 14, 15, 16, 19-23). The capability for representation of stereochemistry is required for any comprehensive treatment of: 1) interpretation of spectroscopic data (see References 15, 16, 19-23); 2) prediction of spectroscopic data (see References 15, 16, 19-23); 3) induction of rules relating known molecular structures to observed chemical or biological properties (see Reference 19,24,26). These elements, taken together, will yield a general system for computer-aided structural analysis (the SASES system) with potential for applications far beyond the specific task of structure elucidation.

Parallel to our program development we have embarked on a concerted effort to extend to the scientific community access to our programs, and critical parts of our research effort are devoted to methods for promoting this resource sharing. Our rationale for this effort is that the techniques must be readily accessible in order to be used, and that development of useful programs can only be accomplished by an extended period of testing and refinement based on results obtained in analysis of a variety of structural problems, analyzed by those scientists actively involved in solutions to those problems. Our efforts in this area are summarized in Section II.A, Scientific Collaboration and Program Dissemination).

B. Medical Relevance and Collaboration

The medical relevance of our research lies in the direct relationship between molecular structure and biological activity. The sciences of chemistry and biochemistry rest on a firm foundation of the past history of well-characterized chemical structures. Indeed, structure elucidation of unknown compounds and the detailed investigation of stereochemical configurations and conformations of known compounds are absolutely essential steps in understanding the physiological role played by structures of demonstrated biological activity. Our research is focussed on providing computational assistance in several areas of structural chemistry and biochemistry, with primary attention directed to those aspects of the problem which are most difficult to solve by strictly manual methods. These aspects include exhaustive and irredundant generation of constitutional isomers, and configurational and conformational stereoisomers under chemical, biological and spectroscopic constraints with a guarantee that no plausible stereoisomer has been overlooked.

Although our programs can be applied to a variety of structural problems, in fact most applications by our group and by our collaborators are in the area of natural products, antibiotics, pheremones and other biomolecules which play important biochemical roles. In discussions of collaborative investigations involved with actual applications of our programs we have always stressed the importance of strong links between the structures under investigation and the importance of such structures to health-related research. This emphasis can be seen by examination of the affiliations of current DENDRAL-related investigators and the brief description of current collaborative efforts in Interactions with the SUMEX-AIM Resource.

C. Highlights of Research Progress

In this section we discuss briefly some major highlights of the past year and research currently in progress.

1. Past Year

1.1 Programs for Interpretation and Prediction of Spectral Data. We are actively pursuing several novel approaches to the automated interpretation of spectral data, concentrating on carbon-13 magnetic resonance (CMR), proton magnetic resonance (PMR) and mass spectral (MS) data. These approaches utilize large data bases of correlations between substructural features of a molecule and spectral signatures of such features. Our approaches are unique in that: 1) we can incorporate stereochemical features of substructures into the data bases; and 2) we can use the same data bases for both interpretation and prediction of data.

For either interpretation or prediction of magnetic resonance data, stereochemical substructure descriptors are absolutely essential. Resonance positions are a strong function of the local environment of a resonating atom, including position in space relative to other neighboring atoms. Descriptors which include the three dimensional relationships among

atoms in a substructure are required in order to obtain meaningful correlations. We now have programs for the interpretation (see References 15,19,21,23), prediction (see references 15,19-21), and assignment (see Reference 22) of carbon-13 nuclear magnetic resonance spectra which make use of an expanded carbon-13 NMR data base. We have completed preliminary work on a program for prediction of proton NMR spectra. All these programs use a structure and substructure representation which incorporates configurational stereochemistry (see reference 16) and make use of data bases.

Such data bases can be used to interpret spectral data to obtain substructures to be used in CONGEN and GENOA, the structure generating programs (see References 15, 17). Continued automation of this aspect of structure elucidation will significantly ease the burden on the structural biochemist because the computer-based files are much more comprehensive and easier to use than correlation tables or diffuse literature sources. The same data bases can be used to predict spectral signatures in the context of a set of complete molecular structures. Comparison of predicted and observed spectra allows a rank-ordering of candidates and will be very useful in directing the attention of the investigator to the most plausible alternatives (see References 7, 8, 15, 20).

1.2 Improvement in the GENOA program for structure generation with overlapping atoms. A significant improvement has been made to experimental versions of GENOA which remove the requirement that a fixed molecular formula be used. This allows a researcher to investigate problems in which this information is unknown. Instead of a fixed composition, a range of compositions is allowed. This program, named RANGEN, will be useful for less specified problems and in particular to the problem of designing molecules (see section 2.1 below).

1.3 Extension of treatment of stereochemistry to include conformations. A computer program has been designed and written which provides a unique (canonical) designation for a conformation of a chemical structure or substructure based on earlier theoretical work (see Reference 24). This program takes as input a structure with 3-dimensional coordinates or cruder conformation designations and gives as output a structure or substructure with a unique conformation designation. Each rotatable bond can have discrete designations which represent a range of torsional angle positions of arbitrary fineness. This program represents a significant step in filling the final "missing link" in our structure representation.

1.4 Molecular Modelling and Graphics. In the past year we have purchased a Megatek Whizzard 7210 vector refresh graphics system and software necessary to interface with our programs and the SUMEX computer system. We have written a program called BUILD3D (see reference 25) which takes as input an augmented (with configurational stereochemistry) connection table representation of a chemical structure and gives as output three-dimensional coordinates. These are converted to a picture on the screen of the Megatek. This picture can be rotated, translated, etc. on the terminal locally with no need of the host computer therefore not adding